Open

INTRODUCTION

quantitative approaches to comparative analyses: data properties and their implications for theory, measurement and modelling

robert neumann[a,*] *and peter graeff*[b]

[a]Technische Universität Dresden, Dresden, 01069, Germany
E-mail: robert.neumann@tu-dresden.de
[b]University of Kiel, Christian-Albrechts-Platz 4, Kiel, 24118, Germany
E-mail: pgraeff@soziologie.uni-kiel.de

*Corresponding author.

doi:10.1057/eps.2015.59

Abstract

While there is an abundant use of macro data in the social sciences, little attention is given to the sources or the construction of these data. Owing to the restricted amount of indices or items, researchers most often apply the 'available data at hand'. Since the opportunities to analyse data are constantly increasing and the availability of macro indicators is improving as well, one may be enticed to incorporate even qualitatively inferior indicators for the sake of statistically significant results. The pitfalls of applying biased indicators or using instruments with unknown methodological characteristics are biased estimates, false statistical inferences and, as one potential consequence, the derivation of misleading policy recommendations. This Special Issue assembles contributions that attempt to stimulate the missing debate about the criteria of assessing aggregate data and their measurement properties for comparative analyses.

Keywords reliability; validity; measurement; quantitative analysis; comparative politics; comparative sociology

(385–393) 1680-4333/15 www.palgrave-journals.com/eps

INTRODUCTION

The social sciences are witnessing an ever increasing supply of data at the aggregate levels on several key dimensions of societal progress or politico-institutional conditions. Next to standardised sources for comparing countries worldwide (Solt, 2014), a bulge of indicators have been introduced over the past three decades to allow for comparative analyses regarding such issues as levels of perceived corruption, quality of governance, environmental sustainability, political rights and democratic freedom. And while there is an abundant use of these macro data, less attention has been given to the sources or to the construction of these data. Despite the spike in data availability, information on countries or regions often remains restricted to only a handful of indicators compiled by organisations that have the resources and know-how to offer worldwide a coverage of countries. Due to this restricted amount of indices or items, researchers for the most part apply the 'available data at hand' with only little consideration of their measurement properties.

There already have been attempts to address questions of data quality within the community of comparative political science. Herrera and Kapur (2007) try to foster the debate about the quality of comparative data sets by highlighting the three components of validity, coverage and accuracy. Mudde and Schedler (2010) discuss the challenges of data choice, distinguishing between procedural and outcome-oriented criteria when data quality is to be assessed. They relate the procedural criterion to aspects of transparency, reliability and replicability of data. The latter criteria is connected to validity, accuracy and precision (Mudde and Schedler, 2010: 411). Both groups of authors agree that research on data properties usually offers little scientific rewards, but that the debate about the measures is crucial and requires constant stimulation.

A few landmark books and articles have laid out some fundamental guidelines and approaches concerning case selection, operationalisation and implications for comparative model testing at the macro level (see for instance King *et al*, 1994; Adcock and Collier, 2001; Gerring, 2001). Yet it appears that the discussion within comparative research about measurement properties of different indicators lags the ongoing application of numerous indices in all sorts of comparative empirical research. That is, theoretical and empirical work with new and improved measurements has so far refrained from the opportunity to enhance an exchange about the conceptual framework for comparative multivariate modelling. Furthermore, it often remains problematic to grasp the core intentions of different streams of knowledge production especially when the computation of new cross-country indices was performed in response to prior criticism of existing measures.

DATA PROPERTIES AND THEIR TRADE-OFF

Judging data properties from a qualitative and quantitative perspective, King *et al* (1994: 63, 97) propose the criteria of unbiasedness, efficiency and consistency. In particular they concentrate on the inferential performance of measures. Here, bias relates to the property to introduce specific variance into the measurement, which in turn leads to non-random variation between different or repeated applications of the measure in inferential tasks. For example, Hawken and Munck (2011: 4) report that ratings on perceived corruption made by commercial risk assessment agencies systematically rate economies as more corrupt than surveys of business executives, representing a bias 'which does not seem consistent with

random measurement error'. Efficiency relates to the variance of a measure when taken as an estimator. The simple idea is that an increase in sample size will likely reduce the variance of a measure and will measure a phenomenon more efficiently. But, even King *et al* (1994: 66) emphasise that these two properties come with a trade-off that is not always easily reconcilable to achieve consistency, most likely in the form that researchers should allow for more bias in their measure if they achieve larger improvements in efficiency. They do not elaborate on consistency further, although they obviously relate it to reliability, which points towards traditional criteria or properties of measurement theory.

This traditional approach of (psychometric) test or measurement theory usually provides social scientists with a framework to think about properties of measures or data. That is, the criteria of validity and reliability remain the cornerstones of any discussions about measurement properties.[1] One can define reliability as an 'agreement between two efforts to measure same trait through maximally similar methods' (Campbell and Fiske, 1959: 83). Usually, this translates to a test of internal consistency of an indicator or test-retest approaches to check whether the systematic variation of an observed phenomenon can be captured by an empirical measure, at several points in time or across different (sub-) samples (Nunnally and Bernstein, 1978: 191). Validity represents a more demanding measurement criterion. A few authors have put forward conceptual approaches to address the problems of constructing indices under the perspective of measurement validity (e.g., Bollen, 1989; Adcock and Collier, 2001). While measurement validity may be broadly defined as the achievement that '… scores (including the results of qualitative classification) meaningfully capture the ideas contained in the corresponding concept' (Adcock and Collier, 2001: 530), it consists of various subcategories such as content, construct, internal/external validity, convergent/discriminant validity and even touches upon more ambitious concepts such as ecological validity as well. These various dimensions also reflect a variety of sources for measurement errors, whether stemming from the process data collection (randomisation versus case selection), survey mode and origin of data, data operationalisation or aggregation of different data sources.

'… the criteria of validity and reliability remain the cornerstones of any discussions about measurement properties'.

Three aspects require us to think harder about the feasibility of these classical concepts of measurement theory. First, the increasing availability of data for the computation or aggregation of macro indicators should improve the reliability of measurements. In fact, it seems that econometricians have completely abandoned the idea of measurement validity and instead focus on statistical techniques for aggregating data. For instance, a recent debate has yielded the impression that reliability remains the main goal to be established, while the concept of validity are not treated as equally important (see the discussion between Kaufmann *et al* (2010) and Thomas (2010)). The problem with the idea to increase the reliability of measures arises at the point when validity is sacrificed due to 'methodological contamination' (Sullivan and Feldman, 1979: 19), especially with regards to the notion that reliability 'represents a *necessary* but not *sufficient* condition for validity' (Nunally and Bernstein, 1978: 192, italics in the original). Hence, aggregated or broadly defined measures that are

unable to discriminate concepts and which are theoretically distinct – and hence are not supposed to be measured by the initial approaches – do not necessarily represent threats to the reliability, but rather to the validity. This is especially the case in empirical tests of theoretical predictions regarding the determinants or consequences of certain politico-institutional conditions, where invalid measures are likely to generate biased coefficients due to measurement error among independent or even dependent variables (Herrera and Kapur, 2007). To this end, results will subsequently lack generalisability. For example, combining several reliable measures of the same phenomena to increase the reliability of the aggregate measure can only claim to be unbiased if all underlying measures capture the same portion of systematic variation in a phenomenon and are able to exclude random measurement error equally well. Testing theories with aggregate measures always comes with the caveat of introducing random measurement error into a measure that is supposed to only represent systematic variation in a phenomenon (see for instance Bollen, 2009 for a discussion), despite being highly reliable.

The potential for a trade-off between reliability and components of validity leads to the second aspect to keep in mind when thinking about measurement properties: Lack of validity may only bother researchers who refer to a *theory-driven approach* of quantitative analyses. The shift towards a *data-driven approach* puts less emphasis on the underlying theory from which one derives hypotheses to be tested. Hypothesis testing may even be the least important aspect of statistical modelling (Varian, 2014: 5). Instead, the goals of data analyses are prediction, forecasting specific behaviours, events or outcomes based on large sets of data, prior knowledge or prior evidence. Due to large amounts of data available and the increasing computer capacities that have enabled the widespread use of Bayesian approaches or machine learning techniques in the social sciences (see Gelman *et al*, 2014; Jackman, 2009), claims can be made that measurement properties that derive their ideas from a theory-driven perspective may lose its relevance. Given this shift, it implies an increasing importance for concepts such as reliability or *predictive* validity that appear closer to the data-driven approach.[2]

The third challenge confronts comparative scholars working with individual-level data. Here, the extension and longevity of survey programmes such as the World Values Surveys or the International Social Science Project (ISSP) have made the application of multilevel models for comparative cross-sectional longitudinal analyses feasible (Beck, 2007; Fairbrother, 2014). Given these opportunities, one core assumption is that measurement invariance holds across countries. That is, questionnaire items capture the same underlying concept across different contexts of data collection in a similar way. On the other hand, the theoretical emphasis on the contextuality of social phenomena creates a desire to reflect such idiosyncratic characteristics of a society within the subsequent measurements approaches.

This creates another trade-off for scholars within the respective research communities. As in the case of reliability and validity, contextually reliable measures can come with a lack of measurement invariance. Given that measurement invariance is tested via its discrepancy to some theoretical model, the shift to data-driven approaches may affect the importance of this particular measurement property in a similar fashion as illustrated for the relationship between reliability and validity.

We perceive this development as neither definitive nor one-dimensional. Measurement theory and the concepts

like validity remain crucial to evaluate and apply the right instruments and to know where to look when research questions are to be answered. That is, how to think or assess the properties of data becomes one crucial aspect of any empirical endeavour. But they seldom represent the only criteria for assessing the characteristics of data. Our own work was concentrated on the aspect of comparing different indices by their measurement properties (Neumann and Graeff, 2010, 2013). One conclusion from this work is that researchers face certain incentives that require decisions on how to cope with the aforementioned trade-offs when measures from comparative data are applied.

THE EDITED SPECIAL ISSUE

Despite the known problems with comparative data, only a few questions remain answered and the stream of new indicators constantly enhances new challenges facing current comparative research. Some key problems can be summarised as follows: How to account for the contextuality of measuring country characteristics while maintaining comparability? What are the consequences when prior knowledge and existing empirical findings are to be included into the derivation of existing and new indicators? How to assess the accuracy of an index and how to even define or measure accuracy in a measurement sense?

This edited issue comprises papers in which the properties of applied aggregate data and the underlying sources for the analysis are explicitly reflected. As the authors bring in different methodological backgrounds, the papers apply the variety of contemporary approaches dealing with reliability and validity. This does not always coincide with a psychometric notion of constructs or measurement criteria. The authors do not, however, fall prey to typical publication strategies such as reporting only significant and/or theoretical congruent results instead of null-results (Gelman and Loken, 2014). All papers share the ambition to accurately reflect the underlying theoretical meaning of the constructs of interest. By this, they refer to the above mentioned key questions in their own way.

Susanne Pickel *et al* (2015) present a new framework for comparative social scientists that tackles one of the most prominent topics in political research: the quality of democracy. In particular, the authors propose a framework to assess the measurement properties of three prominent indices of the quality of democracy. This evaluative process requires both the integration of theoretical considerations about the definitional clarity and validity of the underlying concepts as well as empirical concerns about choice of data sources or procedures of operationalisation and aggregation. Their contribution picks up several important points when one deals with the measurement of macro phenomena. First, although the definition of a concept that encompasses concept validity may vary between researchers or research schools, an assessment of the measurement properties remains tied to rather objective criteria like reliability, transparency, parsimony or replicability. Second, the assessment of a concept and its measurement characteristic ultimately face the challenge of measuring contextual characteristics of a political system as close as possible while adhering to more general measurement principles. The latter represents a task for researchers who want to investigate the comparability of indices. Pickel *et al* apply a framework that includes twenty criteria, focusing on three indices of quality of democracy. The authors state that a theory-based conceptualisation represents the necessary condition for an attempt to face the (potential) trade-off between the adequacy of a measure and its property to

compare it with other measures in a meaningful way.

Mark David Nieman and Jonathan Ring (2015) pick up one of the other big topics of political research: human rights. Their starting point is that all researchers dealing with country data on human rights have to rely on a restricted number of data sources. Namely, the Cingranelli-Richards (CIRI) or the Political Terror Scale (PTS) represents two widely used indices that are both constructed by using the same country reports on human rights violations from the United States State Department and Amnesty International. Their main concern is that if data resources share systematic measurement error, for instance due to politico-ideological or geopolitical bias in the country reports, these properties will likely be reflected in the indices constructed from these data sources. After clarifying why the reports of the US State Department possess such undesirable measurement properties, they propose specific remedies for the problem. Nieman and Ring discuss possible solutions such as data truncation as well as strategies of correcting for systematic bias using an instrumental variable approach. Their replication analysis reveals that the application of the corrected version indeed changes results from prior analyses. Their work highlights the importance of the decisions during the process of indicator choice and subsequent analysis, whereas some choice sets and their consequences regarding inferential reasoning pose conflicting incentives for researchers given the publication bias favouring statistical significant findings (Brodeur *et al*, 2012).

Joakim Kreutz (2015) also scrutinises the methodological foundations of the PTS and CIRI. By referring to both indices, he tries to clarify the connection between human rights and the level of state repression in eighteen West African countries. But instead of focusing on repression levels, Kreutz focuses on changes in repression. By highlighting the importance of repression dynamics, he extends prior evidence on the connection of state repression and politico-institutional factors. From a measurement perspective, disaggregating levels of repression by the direction of change (increase/decrease) and by the nature of repressive actions (indiscriminate, selective targeting) may improve our understanding of the contextual features of repression dynamics. His study provides several implications for current research efforts that try to disentangle the relationship between levels of democracy and state repression.

Alexander Schmotz identifies a gap in the political science literature about the measurement of cooptation, which is the way by which non-members are absorbed by a ruling elite. Concepts of co-optation become particularly important for explaining the upholding of autocratic regimes. As such, issues of co-optation are at the heart of political science research but are only seldom operationalised, especially across time. Schmotz develops an index that is capable to measure several threats to autocratic regimes by social pressure groups. Co-optation is a way to deal with these threats. This topic illustrates some general problems in social science research, namely that theoretical ideas, their predictions about causes and effects, and their testing in empirical research are often intertwined. In such a situation, measurement quality (e.g., content validity) is also related to the performance of the index, in particular if the concept of co-optation refers to a 'seemingly unrelated set of indicators' (Schmotz, 2015). Counterintuitive findings are then of particular importance as in study by Schmotz. He comes up with the conclusion that the concept of co-optation might not be as important as the relevant literature suggests. Such a finding – based on a new index with the potential for testing and improving its measurement features – will incite the discussion in

this field and will most likely lead to refinements of theoretical ideas and their operationalisations.

Barbara Bechter and Bernd Brandl (2015) start with the observation that comparative research is mainly based on aggregates on the national level. This 'methodological nationalism' comes to a dead end if the variance between countries for the variable of interest vanishes (which typically occurs for political regime indicators for western countries, such as the Polity index). They provide an excellent example for an answer to the question about what accounts for the contextuality of comparative research measures as they find that for the field of industrial relations relevant variables reveal more variability across industrial sectors than across countries. This does not imply the meaninglessness of cross-country comparisons. Rather, it opens the perspective to alternative levels of analysis, not only in the field of industrial relations.

William Pollock, Jason Barabas, Jennifer Jerit, Martijn Schoonvelde, Susan Banducci and Daniel Stevens (2015) introduce their study of media effects with the statement that results from analyses of the degree of media exposure on certain attitudes or public opinion are affected by 'data issues related to the number of observations, the timing of the inquiry, and (most importantly) the design choices that lead to alternative counterfactuals' (Pollock *et al*, 2015). In an attempt to provide a comprehensive overview, two identification strategies (difference-in-difference estimator versus within-survey/within-subject) for causal claims from cross- or single country survey data are compared to a traditional approach of statistical inference from regression analyses. Using the European Social Survey and information about media-related events during the data collection process allows them to investigate media effects of political or economic events across countries, across types and number of events as well as across time. With a focus on the external validity of such (quasi-)experimental use of survey data, they are able to generate in parts counterintuitive results regarding the impact of sample size and design effects. Their study emphasises that the process of data collection and design choices have an important impact on subsequent data analyses.

By referring to psychometric techniques, Jan Cieciuch *et al* (2015) raise the question about reliable ways of testing measurement invariance. As a precondition for comparing data, measurement invariance can be determined at the level of theoretical constructs (or latent variables), at the level of relations between the theoretical constructs and their indicators or at the level of indicators themselves. Standard methods to pinpoint measurement invariance based on factor analytical techniques are prone to produce false inferences due to model misspecifications. Cieciuch and his colleagues pick up the discussion in literature about model misspecification and show how one can assess whether a certain level of measurement invariance is obtained. As misspecification must be considered as a matter of degree, their study stimulates the discussion about the question, how much misspecification is acceptable.

Acknowledgements

Parts of this Special Issue follow upon the symposium 'The Quality of Measurement – Validity, Reliability and its Ramifications for Multivariate Modelling in Social Sciences' held at Technische Universität Dresden from 21 to 22 September 2012. Videos of the presentations from the Symposium can be accessed through the website of the symposium at http://tinyurl.com/vwmeasurement. This symposium was financed by the Volkswagen

Foundation, which supported the publication of this special issue as well. We thank all participants of the symposium for their remarks and contributions. Foremost, we thank the Volkswagen Foundation for their financial support.

Notes

1 King *et al* (1994: 25) clarify earlier that the achievement of reliability and validity represent key goals in any social inquiry, whether qualitative or quantitative in nature.

2 This change does not imply a shift from deductive to inductive reasoning from data to theories, because researchers remain bound to deriving their results from a theoretical framework. The nomological core of the data-driven approach stems from the distributive characteristics of different probability distributions. See Gelman and Shalizi (2014) for more details on this line of reasoning.

References

Adcock, R. and Collier, D. (2001) 'Measurement validity: A shared standard for qualitative and quantitative research', *American Political Science Review* 95(3): 529–546.

Beck, N. (2007) 'From statistical nuisances to serious modeling: Changing how we think about the analysis of time-series–cross-section data', *Political Analysis* 15(2): 97–100. doi:10.1093/pan/mpm001.

Bechter, B. and Brandl, B. (2015) 'Measurement and analysis of industrial relations aggregates: What is the relevant unit of analysis in comparative research?' *European Political Science* 14(4): 422–438.

Bollen, K.A. (1989) *Structural Equations with Latent Variables*, New York, NY: Wiley.

Bollen, K.A. (2009) 'Liberal democracy series I, 1972–1988: Definition, measurement, and trajectories', *Electoral Studies* 28(3): 368–374.

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y. (2012) Star wars: The empirics strike back', Paris School of Economics Working Paper 2012–29, pp. 1-.

Campbell, D.T. and Fiske, D.W. (1959) 'Convergent and discriminant validity by the mutitrait-multimethod matrix', *Psychological Bulletin* 56(2): 81–105.

Cieciuch, J., Davidov, E., Oberski, D.L. and Algersheimer, R. (2015) 'Testing for measurement invariance by detecting local misspecification and an illustration across online and paper-and-pencil samples', *European Political Science* 14(4): 521–538.

Fairbrother, M. (2014) 'Two multilevel modeling techniques for analyzing comparative longitudinal survey datasets', *Political Science Research and Methods* 2(1): 119–140.

Gelman, A., Carlin, J., Stern, H., Dunson, D.B., Vehtari, A. and Rubin, D. (2014) *Bayesian Data Analysis*, 3rd edn. London: CRC Press.

Gelman, A. and Shalizi, C. (2014) 'Philosophy and the practice of Bayesian statistics', *British Journal of Mathematical and Statistical Psychology* 66(1): 8–38.

Gelman, A. and Loken, E. (2014) 'The statistical crisis in science data-dependent analysis – a 'garden of forking paths' – explains why many statistically significant comparisons don't hold up', *American Scientist* 102(6): 460. doi:10.1511/2014.111.460.

Gerring, J. (2001) *Social Science Methodology: A Criterial Framework*, Cambridge: Cambridge University Press.

Hawken, A. and Munck, G.L. (2011) 'Does the evaluator make a difference? Measurement validity in corruption research', *Measurement Validity in Corruption Research.*

Herrera, Y.M. and Kapur, D. (2007) 'Improving data quality: Actors, incentives, and capabilities', *Political Analysis* 15(4): 365–386.

Jackman, S. (2009) *Bayesian Analysis for the Social Sciences*, New York: John Wiley & Sons.

Kaufmann, D., Kraay, A. and Mastruzzi, M. (2010) 'Response to 'what do the worldwide governance indicators measure?'', *European Journal of Development Research* 22(1): 55–58.

King, G., Keohane, R.O. and Verba, S. (1994) *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton, NJ: Princeton University Press.

Kreutz, J. (2015) 'Separating dirty war from dirty peace: Revisiting the conceptualization of state repression in quantitative data', *European Political Science* 14(4): 458–472.

Mudde, C. and Schedler, A. (2010) 'Introduction: Rational data choice', *Political Research Quarterly* 63(2): 410–416.
Neumann, R. and Graeff, P. (2010) 'A multitrait-multimethod approach to pinpoint the validity of aggregated governance indicators', *Quality & Quantity* 44(5): 849–864.
Neumann, R. and Graeff, P. (2013) 'Method bias in comparative research: Problems of construct validity as exemplified by the measurement of ethnic diversity', *Journal of Mathematical Sociology* 37(2): 85–112.
Nieman, M.D. and Ring, J.J. (2015) 'The construction of human rights: Accounting for systematic bias in common human rights measures', *European Political Science* 14(4): 473–495.
Nunally, J.C. and Bernstein, I.H. (1978) *Psychometric Theory*, New York: McGraw-Hill.
Pickel, S., Stark, T. and Breustedt, W. (2015) 'Assessing the quality of quality measures of democracy: a theoretical framework and its empirical application', *European Political Science* 14(4): 496–520.
Pollock, W., Barabas, J., Jerit, J., Schoonvelde, M., Banducci, S. and Stevens, D. (2015) 'Studying media events in the European social surveys across research designs, countries, time, issues, and outcomes', *European Political Science* 14(4): 394–421.
Schmotz, A. (2015) 'Vulnerability and compensation – Constructing an index of co-optation in autocratic regimes', *European Political Science* 14(4): 439–457.
Solt, F. (2014) 'The Standardized World Income Inequality Database', Working paper. SWIID Version 5.0, October 2014. http://myweb.uiowa.edu/fsolt/index.html.
Sullivan, J.L. and Feldman, S. (1979) 'Multiple indicators – An introduction' Sage University Paper series in Quantitative Applications in the Social Sciences No. 07–15, Beverly Hills and London: Sage.
Thomas, M. (2010) 'What do the worldwide governance indicators measure?' *European Journal of Development Research* 22(1): 31–54.
Varian, H.R. (2014) 'Big data: New tricks for econometrics', *The Journal of Economic Perspectives* 28(2): 3–28.

About the Authors

Robert Neumann born 1980, earned his Ph.D. in Sociology in 2012 from Technische Universität Dresden, where he also received his Diploma in Economics in 2007. He has worked as Chair of Macrosociology from 2008 to 2012 and currently is working at the chair for Methods of Empirical Social Sciences at Technische Universität Dresden. His research interests focus on quantitative macro research, rational choice theory and survey research methods. Recent publications have appeared in, for example, the *European Journal of Political Research*, the *Journal of Mathematical Sociology* and *Quality & Quantity.*

Peter Graeff is Professor at the Faculty of Faculty of Business, Economic and Social Science at Kiel University. Prior coming to Kiel, he was an assistant professor at Goethe University, Frankfurt, and Düsseldorf University. He received a diploma in economics and a Ph.D. in psychology from Bonn University. His research focusses on empirical research methods and social capital research in its positive (e.g., social trust) and its negative sense (e.g., corruption). Recent publications appeared among others in *Plos one, European Sociological Review, Quality & Quantity*, and the *Journal of Mathematical Sociology*.